Human Tracking Using Particle Filter with Reliable Appearance Model

SICE Annual Conference 2013
September 14-17, 2013, Nagoya University, Nagoya, Japan

Human Tracking using Particle Filter with Reliable Appearance Model

Sangeun Lee[1] and Keiichi Horio[1,2]

[1]Graduate School of Life Science and Systems Engineering,
Kyushu Institute of Technology, Kitakyushu, Japan
(Tel: +81-93-695-6127; E-mail: lee-sangeun@edu.brain.kyutech.ac.jp, horio@brain.kyutech.ac.jp)
[2]Fuzzy Logic Systems Institute, Iizuka, Japan
(Tel: +81-948-24-2771; E-mail: horio@brain.kyutech.ac.jp)

Abstract: In this paper, we present a human tracking algorithm that can work robustly in complex environments such that serious occlusion, various appearances and abrupt motion changes occur in the scenario. Our tracking framework is well known particle filter based on Condensation algorithm. In the observation model of the particle filter, we establish RAM(Reliable Appearance Model) which exhibits high discriminative performance in particular for human tracking. The RAM is to describe a target as features from local descriptors. In order to extract practical features from a larger number of local descriptors for robust tracking, the features were employed by boosting algorithm. The components of the features are utilized color and shape based-models. Experimental results demonstrate that our approach tracks the target accurately and reliably when position and scale are changing as well as occurrence of occlusion.

Keywords: Human tracking, Reliable Appearance Model (RAM), Particle filter, Local descriptor, Feature extraction

1. INTRODUCTION

Visual tracking of a target in real scene is one of the most important topics in the field of computer vision. It has been widely applied to such as surveillance, robotics, and human computer interaction system[1]. However, object tracking typically contains many uncertainties of information such as various pose changes, significant occlusions and measurement noise in the scenario, the tracking is difficult as well as challenging problem. Especially, human tracking rather than other objects contains abrupt changes in the appearance and the poses of the target. Fig.1 shows examples of the human tracking. In order to cope with those difficulties, approaches of particle filter framework have been widely adopted as an effective method of the human tracking. The particle filter is suitable for tracking due to the approximation of probability distribution and requiring only anterior information. Relying on probability distribution in video sequences, considering the information for future frames, the problem of frequent occlusions of a target can be solved effectively. A key element of particle filter in video sequences is a selection of image feature which describe closely a target as an appearance model[2, 5, 8, 14] such as color, edge and so on. The image features are affected by not only lighting change but also pose changing of the target. The color-based image features such as histograms in RGB or HSV have advantages for tracking non-rigid and fast moving objects as they are robust partial occlusion, rotation and scaling[7, 9, 10, 13]. The edge-based image feature histogram of oriented gradients (HOG)[3] were reported in particularly robust for human detection. Although HOG descriptor operates on localized cell and upholds invariance to geometric and photometric transformations. It requires high computational costs.

Several approaches for visual tracking have been addressed, we only introduce in detail the works that are design of observation model. Kwon et al.[19] showed the robustness to pose variation, occlusion, significant illumination change and abrupt motion change in real environments by decomposition of observation model. However, this method employed the template image fed into overall pixels of the target image, high computational costs and noise from the background of the target image are concerned. Dominik et al.[2] present adaptive visual tracking in real time. Although the observation model which locally boosted weak classifiers on counter-surround features with RGB is successfully service for tracking the target, they only take the color information into consideration, adding shape information of features is rather than robustly tracking the target. On the other hand, the observation model can effectively solve these concerning identified features to the target, extracted from localized descriptors with the color and shape information.

Fig. 1 Example of human tracking in complex environment. These scenarios include abrupt changes, a variety of poses, occlusion and lighting changes[18, 19].

In this paper, we propose a particle filter framework with appearance-based human identification. The particle filter is based on Condensation algorithm[10] developed in the computer vision community and typically used. As the proposed method, the extracted features

PR0001/13/0000-1418 ¥400

Algorithm 1 Particle filter

1. Initialization $t = 0$
   for $i = 1, \ldots, N$
   sample $X_0^{(i)} \sim p(X_0)$; $t := 1$
   endfor
2. Sampling step
   for $i = 1, \ldots, N$
   sample $\tilde{X}_t^{(i)} \sim P(X_t | X_{t-1}^{(i)})$
   $\tilde{X}_{0:t}^{(i)} = \left(X_{0:t-1}^{(i)}, \tilde{X}_t^{(i)}\right)$
   endfor
   for $i = 1, \ldots, N$
   evaluate importance weight $\tilde{w}_t^{(i)} = \left(Y_t, \tilde{X}_t^{(i)}\right)$
   endfor
   normalize the importance weight
3. Selection step
   for $i = 1, \ldots, N$
   draw N with probability proportional to $\tilde{w}_t^{(i)}$
   add $\tilde{X}_t^{(i)}$ to $X_t^{(i)}$
   endfor
   set $t \leftarrow t + 1$ and go to step 2

Fig. 2 Algorithm of particle filter.

from a target of human is named as a Reliable Appearance Model (RAM). RAM effectively incorporates color and edge information into discriminative features. As color-based image feature, Hue of HSV model that is invariant to lighting variations is adopted. Although it requires heave computational costs, we have a strategy to reduce the computation in which only used pre-defined local image descriptors[12, 15]. A set of most discriminative features is automatically selected from a large number of local image descriptors. This set serves as the features for the observation model of phase in particle filter. In this phase, similarity between query images of each particle and memorized image is calculated based on discriminative features with standard measurements such as Bhattacharyya distance. This strategy performs high discrimination ability with fulfilling requirements of computational costs.

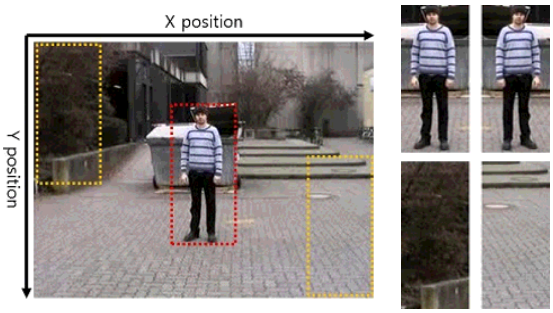
This paper is organized as follows. In Section 2 we briefly present a probabilistic tracking algorithm with a particle filter. In Section 3 our approach which exploits the features for robust object tracking and integrates it into the observation model is explained. In section 4 we describe experiments that the proposed approach is applicable to a wide variety of tracking scenarios In Section 5, we conclude the paper with discussions.

2. PARTICLE FILTER

The particle filter is a technique of estimating of feature state that is based on recursively computing Bayesian formulation. The Bayesian formulation predicts the sequence of hidden parameters based on only the observed data.

2.1 Bayesian formulation

The visual tracking problem is cast an inference task in a Markov model with hidden state variables[16]. The



(a) (b)

Fig. 3 Definition of a target, a positive image and negative images. The initial frame(a) provides a target(b. left top), a positive(b. right top) image, and negative(b. bottom) images.

state variables X_t describes target configuration at time t, and observation Y_t extracted from images. We aim to estimate the value of the hidden state variable of X_t based on given all observation $Y_{1:t} = (Y_1, \ldots, Y_t)$ up to time t. The posterior probability $p(X_t|Y_{1:t})$ is estimated with the following Bayesian formulation. Bayesian estimation updates the posterior probability with the following rule

$$p(X_t|Y_{1:t}) \propto p(Y_t|X_t) \int p(X_t|X_{t-1}) p(X_{t-1}|Y_{1:t-1}) dX_{t-1}, \quad (1)$$

where, $p(Y_t|X_t)$ donates the observation model that measures how well the observation fits the predictions, and $p(X_t|X_{t-1})$ represents the motion model that proposes the next state X_t based on the previous state X_{t-1}. As shown in Fig.2, Condensation algorithm[10] which is based on factored sampling, approximates an arbitrary distribution of observation with a generated set of weighted samples. The distribution is represented by a set of particle $\{(X_t^{(i)}, w_t^{(i)}), i = 1, \ldots, N\}$, where, $X_t^{(i)}$ and $w_t^{(i)}$ denote a state and an associated weight of the i-th particle at time t, respectively. Using importance sampling, Eq. (1) recursively approximated by

$$p(X_t|Y_{1:t}) \approx p(Y_t|X_t) \sum_i w_{t-1}^{(i)} p(X_t|X_{t-1}). \quad (2)$$

Given a set of particles from the previous time $\{(X_{t-1}^{(i)}, w_{t-1}^{(i)})\}$ configurations at the current time $X_t^{(i)}$, are drawn from a proposal distribution

$$q(X_t) = \sum_i w_{t-1}^{(i)} p(X_t|X_{t-1}^{(i)}). \quad (3)$$

The weights are then updated as $w_t^{(i)} \propto p(Y_t|X_t^{(i)})$. The motion model predicts the particle position. The estimate of state of particles is based on a linear extrapolation of the previous state plus Gaussian noise. The current state of the target is estimated as a weighted average over the states of the particles. The observation model evaluates the weight by likelihood, and is generally established the appearance of target from the image that is utilized color, edges or texture as a feature[1].

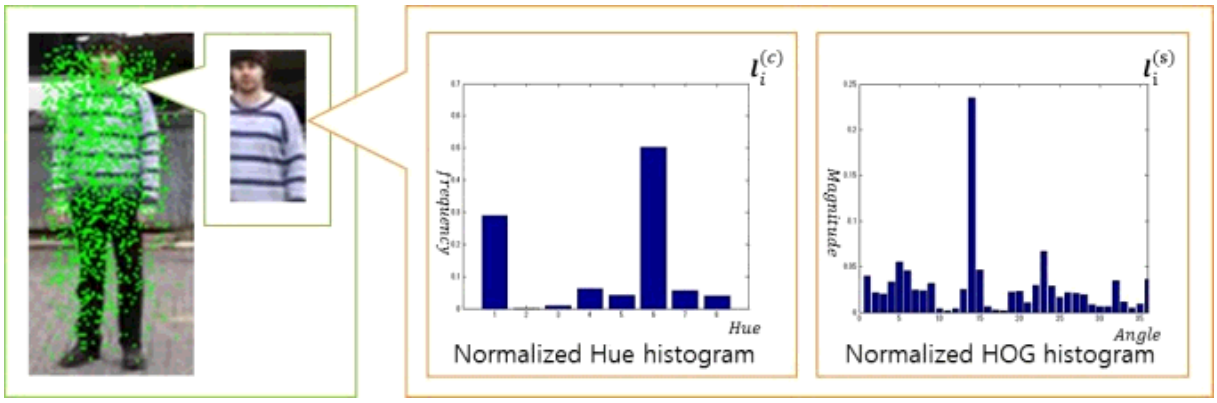


Fig. 4 Description of the local descriptors. A rectangular contains color and shape information.

3. RELIABLE APPEARANCE MODEL

The appearance model is a core in features selection and observation model in the particle filter. We seek RAM which exhibits high discriminative performance so that we extract several descriptors in a large number of local image descriptors by boosting algorithm.

3.1 Definition of learning images

The positive and negative images are required for learning in Boosting algorithm. Thus both of images are defined from the first frame in sequential images as shown in Fig.3. Firstly, the rectangle containing a target of human is manually marked from the frame, and it is considered as target image. A positive image is mirror-reversed image of the target image. The negative images are sampled by sliding window in background in the frame, because there is spares possibility of similar appearance as the target image in background image.

3.2 Local image descriptors

As shown in Fig.4, given an image sample I, a feature of the image I, which is a set of local descriptors, is extracted in the following manner. At first 1800 points are defined in the image I. At point j, a local image around the point is extracted, here, the size of the local image is 10, 15 or 20 in width, and height is the same, half or twice of the width. For each local image, a normalized histogram of hue in HSV color space $\boldsymbol{l}_j^{(c)}$ and a normalized HOG histogram $\boldsymbol{l}_j^{(s)}$ are calculated. The detail about the features is explained in the following. As a result, 3600 normalized histograms can be obtained. The set of these histograms $\{\boldsymbol{l}_j^{(c)}, \boldsymbol{l}_j^{(s)} | j = 1, ..., 1800\}$ is used as the feature of the image I.

3.2.1 Color-based model

In the tracking algorithm in computer vision, color information is widely used to calculate the similarity between two images. The histograms which are characterized as color distribution in the region achieve robust against to non-rigidity and rotation. The hue of the HSV color model is chosen in this paper, because the hue is not sensitive lighting condition compared to RGB[6][9]. Furthermore, the circular character of the range can be represented from the $360°$ range of possible combinations of the $redness$, $greenness$ and $yellowness$. Thus the hue information has the benefit of computational costs. In our approach, the color distributions of the hue are discretized into a histogram of 8 bins over the region from 8 dimensional vectors $\boldsymbol{l}_j^{(c)} = (l_{j1}^{(c)}, \cdots, l_{j8}^{(c)})$. In the simulation, the histogram is normalized.

3.2.2 Shape-based model

The shape information in the image is helpful as well as color information to represent the geometric information such as human. HOG descriptors[3] presents a human detection algorithm with excellent detection result. Their method uses a dense grid of histogram of oriented gradient. Each detection window is divided into cells of size 8×8 pixels and each group of 2×2 cells is integrated into a block in a sliding fashion, so blocks overlap with each other. Each cell consists of a 9-bin histogram of oriented gradients and each block contains a concatenated vector of all its cells. Each block is thus represented by a 36 feature vector that is normalized to a unit length. In this paper, as a shape descriptor $\boldsymbol{l}_j^{(s)}$, HOG descriptor is adopted. A region where, is divided 2×2 cells with 9-bin. The local descriptor contains 36 dimensional vectors $\boldsymbol{l}_j^{(s)} = (l_{j1}^{(s)}, \cdots, l_{j36}^{(s)})$, and represents a normalized local descriptor.

3.3 Descriptor selection

How to effectively select reliable local descriptors is a key problem for robust performance in the tracking. The valuable features are extracted from a large number of the local descriptors, *i.e.* a set of local feature $\{\boldsymbol{l}_j^{(c)}, \boldsymbol{l}_j^{(s)} | j = 1, ..., 1800\}$. Given the color and shape of local descriptors, Bhattacharyya[6] can be used to compute a similarity score between two images. Although the negative images have low similarity, the positive image has a high similarity against to the target image. Since the human has a constancy appearance such shoulder and head, this approach is suitable. The both similarity scores are required for learning data by boosting. In this paper, Real Adaboost[11] is adopted. The algorithm of real ad-

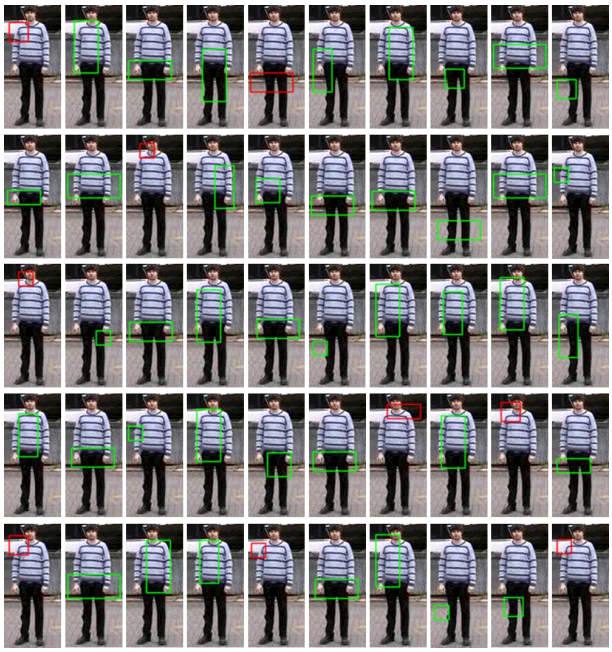


Fig. 5 A set of selected features in the experiment concerning the scenario B. The weak classifiers are selected by Adaboost. The local descriptors of Hue and HOG are indicated by red and green, respectively.

aboost widely used is as follows:

$$h(\cdot) = \text{sign}\left[\sum_{n=1}^{N} h_n(\cdot) - \lambda\right], \tag{4}$$

where, $h(\cdot)$ and $h_n(\cdot)$ are the outputs of classifier and n-th weak classifier, respectively. Although the general algorithm combines with weak classifiers for the same input, we use the weak classifier independently of each feature. This strategy has the advantage of the computational cost and accuracy[12]. The selected local descriptors are represented as the set of the features and are called RAM. Moreover, in calculation of a likelihood in observation model, these features are required usefully.

4. EXPERIMENTS AND RESULTS

4.1 Features performance

The selected features by the proposed method are required to qualify as the features. To demonstrate the performance of the selected features, the discriminate simulation, in which the target of human is identified from images with other human or without human, was achieved. As a learning set for the feature selection, the image shown in Fig.3 is used. Fig.5 shows 50 selected features. In Fig.5, red and green rectangles represent selected features of color and shape information, respectively. Classification ability for training data was 100%.

As a testing data, we use CU-dataset[20]. In the data set, 2053 positive images and 6253 negative images are included. The purpose of this study is to track the target of human, and all images in CU-dataset should be identified that the target human is not included. As the results, all images were correctly classified to the images without the target of human. This result indicates that the selected features contain characteristic information which is suitable for identifying the target of human.

Table 1 Configuration of datasets

Configuration		A	B	C	D	E	F	G	H	I
Camera	Fixed	○	-	-	-	-	-	-	○	-
	Move	-	○	○	○	○	○	○	-	○
Occlusion	Occur	○	-	○	○	○	○	○	○	-
	Not occur	-	○	-	-	-	-	-	-	○
Lighting	Change	-	-	-	-	-	-	-	-	○
	Not change	○	○	○	○	○	○	○	○	-
Pose	Change	-	-	○	-	-	○	○	○	-
	Not change	○	○	-	○	○	-	-	-	○

Table 2 Number of selected features for the scenarios.

	A	B	C	C	E	F	G	H	I
HUE	35	41	38	21	41	8	16	44	37
HOG	15	9	12	29	9	42	34	6	13

4.2 Evaluation of accuracy of tracking

To evaluate the accuracy of tracking results, we measure the recall ρ , the precision ν and the *F-measure*[17, 18]. These measures are calculated by

$$\rho = \frac{A_t^E \bigcap A_t^G}{A_t^G}, \nu = \frac{A_t^E \bigcap A_t^G}{A_t^E} \tag{5}$$

where, A_t^E and A_t^G represent the estimated area and ground-truth at time t, respectively. Both of the recall and the precision should have high values for good tracking quality. Combined these values can be used to define the tracking quality by the *F-measure*.

$$F = \frac{2\nu\rho}{\nu + \rho}, \tag{6}$$

where, when the ground-truth and estimated area are perfectly overlapped, the *F-measure* becomes 1.0. If *F-measure* is higher than 0.5, the target can be considered as correctly tracked at that time.

4.3 Tracking results

This section reports tested 9 video sequences ranging from surveillance robot vision and dynamic sports scene[18-21]. As shown in Table 1, these datasets are categorized according to environmental considerations such as condition of the camera, occlusion, lighting and pose change. Table 2 shows the numbers of selected features corresponding to the scenarios. The recall, precision and *F-measure* shown in Table 3 are calculated between the ground-truth state and the estimated state.

4.3.1 Scenario A: static camera with occlusion and pose change

The camera is stationary, and two humans walk front of a target and then the target human walks to sideways. This experiment is assumed as a situation of surveillance. As shown in Fig.6, the target human is perfectly tracked as well as an *F-measure* is over 0.9.

Table 3 Accuracy of human tracking corresponding to the 9 scenarios.

	A	B	C	D	E	F	G	H	I
Recall	0.941	0.977	0.973	0.818	0.981	0.940	0.950	0.864	0.913
Precision	0.916	0.561	0.745	0.859	0.679	0.515	0.596	0.763	0.334
F-measure	0.926	0.707	0.833	0.822	0.794	0.661	0.725	0.805	0.436

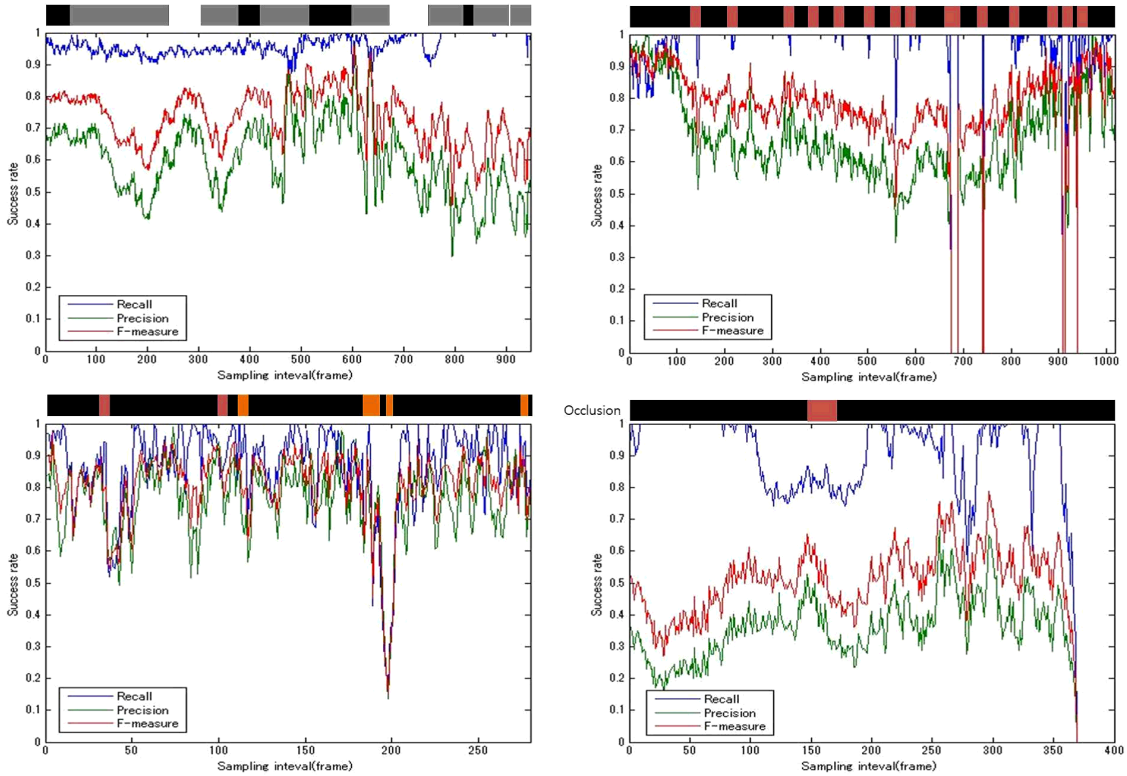


Fig. 6 The results of experiments of B, E, H and I in frame. The white, Gary, orange and red in upper of graphs are indicated appearance such as sideways, back, overlap and occluded target respectively.

4.3.2 Scenario B: non-rigid object in an outdoor scene

In our strategy, the features just required from initial frame. Thus, it can doubt to handle the variety of appearance of target poses. Since this scenario contains various poses such as turning around multiple times. As shown in Fig.5, features are selected from the initial frame. The number of color and shape features are 9 and 41, respectively. Interestingly, the shape features are distributed evenly around the target silhouette. On the other hand, color features are relatively selected in skin color such as the face. The *F-measure* over frames in Fig.6 indicates that our approach successfully tracks the target human in spite of various appearances such as sideways and back.

4.3.3 Scenario C: partial occlusion with moving camera

The camera moves so that the human gets half articulated and then visible again. A target human is stationary. The discriminative features are selected from the various positions in the image, then partial occlusion is also naturally overcome. As shown in Table 3, the recall and the *F-measure* are very stable in partial occlusion but the precision has low value. This is because, the estimated area included invisible area unlike supported ground-truth.

4.3.4 Scenario D: full occlusion of a Non-rigid object

The target of human walks along a corridor and becomes fully occluded by a pillar three times and a human once. While the object is not visible temporally, distributed particles by Gaussian noise make rapidly assign weight to particles when the target human is visible again. Thus, RAM is a suitable for full occlusions as well.

4.3.5 Scenario E: following human with occlusion

Since the experimental environment is following a human with moving camera in outdoor, we assumed that the camera is mounted on a mobile robot such as robot vision. In this scenario, other humans occlude the target human 13 times, furthermore color of clothes of some human is similar to that of the target. In the proposed method, not only the color information but also shape information are included as the features, then the *F-measure* is high as shown in Table 3.

4.3.6 Scenarios F and G: crossing a target in an indoor scene

Two humans walk forward in narrow corridors. The target are fully occluded each other two times. The tracking results indicate low *F-measure* value compared to

other experiments. As shown in Table 2, a small number of HOG features are selected. As these results, we figure out that lack of the HOG features certainly affect significant to track human.

4.3.7 Scenario H: real-world with rapid moves

In a real scenario, the target rapidly moves over one or more small temporal intervals. Thus, we experiment with a badminton scenario which consists of abrupt pose change and occlusion. Since the target moves rapidly, this experiment required large distribution of particles compared to other experiments by Gaussian noise. Although the particles rarely move to a teammate temporally while the target occluded by the teammate, the particles rectify themselves.

4.3.8 Scenario I: real-world with lighting changes

The sequence contains the serious lighting change, occlusion and pose changes in skating scenario. This sequence can verify the effectiveness of the features which are combined of shape and color information. As shown in Fig.6, even though the target is successfully tracked against to lighting changes and occlusion, our tracking algorithm is unfortunately failed when the change lighting and the pose occurs at once. This result can be considered that the algorithm is required corresponding to the charge of poses by an adaptive model.

5. CONCLUSION

In this paper, we presented a particle filter framework for human tracking using reliable local descriptor which consists of appearance models with color and shape. The features, which are selected from the large number of local descriptors by boosting algorithm, are suitable for tracking targets using particle filter. Especially, human tracking in various scenes was successfully achieved by combining hue and HOG as the features. Tracking failures caused by variation of target in size suggest as future work. It required a further refinement in particle filter. Additional future work could explore that our tracking algorithm is applied to real time such as mobile platform.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey", *ACM Computing Surveys (CSUR)*, Vol.38, No.4, 2006.

[2] D.A. Klein, D. Schulz, S. Frintrop and A.B. Cremers, "Adaptive Real-Time Video-Tracking for Arbitrary Objects", *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp.772-777, 2010.

[3] N. Dalal and B. Triggs., "Histograms of Oriented Gradients for Human Detection", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.886-893, 2005.

[4] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier and L.V. Gool., "Robust Tracking-by-Detection Using a Detector Confidence Particle Filter", *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pp.1515-1522, 2009.

[5] K. Smith, D.G. Perez and J. Odobez, "Using Particles to Track Varying Numbers of Interacting People", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.962-969, 2005.

[6] K. Nummiaro, E. Koller-Meier and L. Van Gool., "Object Tracking with an Adaptive Color-Based Particle Filter", *Proc. Symp. on Pattern Recognit. DAGM*, pp.353-360, 2002.

[7] C.H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.1217-1224, 2011.

[8] P. Perez, C. Hue, J. Vermaak and M. Gangnet, "Color-Based Probabilistic Tracking", *Proc. European Conf. on Computer Vision*, pp.661-675, 2002.

[9] T. Tung and T. Matsuyama, "Human motion tracking using a color-based particle filter driven by optical flow", *Proc. 1st Int. Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08*, 2008.

[10] M. Isard and A. Blake, "Condensation-conditional density propagation for visual tracking", *Int. Journal of Computer Vision*, Vol.29, No.1, pp.5-28 1998.

[11] R.E. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", *Machine Learning*, No.37, pp.297-336, 1999.

[12] W. Nam, B. Han and J.H. Han, "Improving Object Localization Using Macro feature Layout Selection", *Proc. IEEE Int. Workshop on Visual Surveillance*, pp.1801-1808, 2011.

[13] W. Nam, B. Han and J.H. Han, "Real-time tracking of non-rigid objects using mean shift", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vol.2, pp.142-149, 2000.

[14] S. Zhou, R. Chellappa and B. Mogghaddam, "Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters", *IEEE Trans. Image Process*, pp.1491-1506, 2004.

[15] J. Kwon and K.M. Lee, "Tracking of a Non-Rigid Object via Patch-based Dynamic Appearance Modeling and Adaptive Basin Hopping Monte Carlo Sampling", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.1208-1215, 2009.

[16] D.A. Ross, J. Lim, R. Lin and M. Yang, "Incremental learning for robust visual tracking", *Int. J. Computer Vision*, Vol.77 pp.125-141, 2008.

[17] K. Smith, D.G. Perez and J. Odobez, "Using Particles to Track Varying Numbers of Interacting People", *Proc. IEEE Int. Conf. on Computer Vision and Computer Recognition (CVPR)*, Vol.1 pp.962-969, 2005.

[18] J. Kwon and K.M. Lee, "Wang-Landau Monte Carlo-Based Tracking Methods for Abrupt Motions", *IEEE Trans. Pattern Anal. Mach. Intell.*, pp.1011-1024, 2013.

[19] J. Kwon and K.M. Lee, "Visual tracking decom-

Fig. 7 Tracking results in the down-sampled all of experiments from A to I.

position", *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.1269-1276, 2010.

[20] CU-dataset, "http://www.vision.cs.chubu.ac.jp/JointHOG/"

[21] BoBot (Bonn Benchmark on Tracking) dataset, "http://www.iai.uni-bonn.de/ kleind/tracking/"